Supplementary Methods

Genome variation and evolution of the malaria parasite *Plasmodium falciparum*.
Daniel C. Jeffares, Arnab Pain, Andrew Berry, Anthony V. Cox, James Stalker, Catherine E. Ingle, Alan Thomas, Michael A. Quail, Kyle Siebenthall, Anne-Catrin Uhlemann, Sue Kyes, Sanjeev Krishna, Chris Newbold, Emmanouil T. Dermitzakis, Mathew Berriman.

Collection of PFCLIN sample and DNA extraction

The Ghanaian clinical isolate (PFCLIN) was obtained by erythrocytapheresis used as adjunct treatment to quinine to treat hyperparasitaemia in a 31 year old woman (see [1]). The patient had returned from a holiday in Ghana 2 weeks before admission to the Intensive Care Unit at St. George's Hospital. She presented with a four day history of nausea, vomiting and diarrhoea and three days of rigors. Although of Ghanaian origin, she was born in the United Kingdom and there was no history of previous episodes of malaria. She had discontinued doxycycline antimalarial prophylaxis because of indigestion and concerns about phototoxocity. On admission, she was febrile (39°C oral temperature) with a heart rate of 98 beats per minute and a blood pressure of 105/90 mm Hg (supine). She was fully orientated with no other abnormalities on examination. Her peripheral blood film confirmed *Plasmodium falciparum* infection (26% parasitaemia). Her haemoglobin was 117 g/L (normal 120-160 g/L), white cell count was 10.2 x 10^9/L (4-11 x 10^9/L), and platelet count was 53 x 10^9/L (150-450 x 10^9/L). Her creatinine was 87 μmol/L (60-110 μmol/L), total bilirubin was 64 μmol/L (0-17 μmol/L) and ALT 23 U/L (0-40 U/L), albumin 42 g/L (35-48 g/L). She was treated with a loading dose of quinine (20 mg/kg salt) followed by conventional maintenance doses and was discharged after an uncomplicated course, when her parasitaemia had cleared.

Molecular studies were approved by the Wandsworth Local Research Ethics Committee (UK). About 400 ml of enriched infected red blood cells (~ 70% parasitemia after erythrocytapheresis) were available for harvesting of parasite DNA. The PFCLIN isolate

was not culture-adapted. Red blood cell membranes were lysed in Erythrocyte Lysis buffer (Qiagen) and samples were aliquoted and stored in Trizol at -70°C. Approximately 20 ml of thawed cells were processed for DNA, using proteinase K digestion and phenol/chloroform extraction. The resulting DNA yield (3.5 mg) was approximately twice that expected from the total input number of parasitized cells, so we assessed the level of human DNA contamination. Duplicate Southern blots of EcoRI digested DNA were prepared, comparing 2 μg of the PFCLIN DNA sample to a titration series of human and parasite genomic DNA, mixed at known ratios. One blot was hybridized with a total human genomic DNA probe (at 65 °C), the other with a total parasite genomic DNA probe (at 50ºC; hybridization, blot washing and autoradiography as previously described[2]). Probes were prepared from 50 ng total genomic DNA, labelled with alpha 32P-dATP according to manufacturer's instructions (Megaprime, Amersham/GE). From the films of these blots, we estimated that the PFCLIN DNA sample was approximately 3:1 human: parasite DNA, and that a further purification step would be required to remove the human DNA component. Human DNA was then removed by two rounds of centrifugation in a CsCl gradient in the presence of Hoechst-33258 (Ref [3])**.**

P. reichenowi sample. Prior to commencement of the study was approved by the Institutional Ethics Committee Dierexperimentcommissie, DEC) of the Biomedical Primate Research Centre (BPRC) according to Dutch law. In January of 2001, blood was collected from chimpanzee Dennis that had been infected with *P. reichenowi* Oscar strain. 12 days after infection, when the parasitemia was 0.5%, blood was collected and filtered to reduce white blood cell numbers (Plasmodipur, Netherlands). The infection was subsequently treated with chloroquine and resolved fully. DNA was isolated from infected erythrocytes using the PureGene gDNA isolation kit (Gentra systems, Minneapolis. USA) according to manufacturers instructions and subsequently stored at 4 °C.

Read alignment and identification of differences

Paired shotgun clone reads were aligned to the completed *P. falciparum* 3D7 genome[4] (3D7.version2.0, ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/) with SSAHA2 http://www.sanger.ac.uk/Software/analysis/SSAHA2/ (Ref [5]). Only alignments that mapped to a single location on the reference genome and were opposed by their read pair were used in further analysis (single location paired reads). Single nucleotide differences (SNPs), and polynucleotide differences (POLYs, insertion/deletion events, differences of two or nucleotides within five nucleotides of each other, and more complex polynucleotide differences) were identified from these alignments using neighbourhood quality standard[6].

Data quality filters

Many multiple-location and non-opposed reads mapped to the repetitive *VAR* gene clusters[4]. To exclude SNP calls of uncertain location from all such regions, for each project (PFCLIN, IT, *P. reichenowi*) we defined 'uniqueness' metric (U) for each 10 kb block of the genome, where U = [(single location paired reads mapped to block)/(all reads mapped to the block)]. Regions with $U \geq 0.5$ were defined as the 'unique' portion of the genome. Any reads/SNPs/POLYs overlapping non-unique regions were excluded from further analysis.

SNP locations with multiple read coverage were used to determine the cumulative phred score for the SNP from all reads ($\sum P_{SNP}$), and the cumulative phred score for the reference ($\sum P_{ref}$). A SNP was considered validated if $\sum P_{SNP} > \sum P_{ref}$. For each project the minimum validated score (MVS) was determined as the minimum base phred score for which 90% of SNP calls were validated. For PFCLIN and IT the MVS = 42, for *P. reichenowi* the MVS = 24. All subsequent analysis used only SNPs and 'multiple SNP' POLYs with a phred score ≥ MVS or a cumulative phred score (from all reads covering the location) greater than 60 ($1/10^6$ probability of error). All indel POLYs were retained. Conflicting SNP were resolved to the base with the highest cumulative phred score. POLY calls were resolved to remove identical POLYs, but retained overlapping non-

identical calls. Estimates of the total rate of genetic change, dN/dS and MK tests were calculated from a predicted variant sequence generated using SNP and multiple SNP POLYs with SSAHA2 scores ≥ the MVS.

We manually scrutinised the SNPs called from PFCLIN in 12 genes (44124 nt in length). All sequence reads from PFCLIN covering these genes, and further reads were generated by sequencing PCR products obtained from the initial PFCLIN DNA extraction, were aligned to the reference sequence using Gap4 v4.10b.3 (http://staden.sourceforge.net/staden_home.html). SNPs were then manually scrutinised by examining the electrophoretograms. These regions contained 50 SSAHA-identified SNPs that satisfied our MVS quality criteria: 48 were located and verified as correct, including one heterozygous SNP. One was identified but not found to be a SNP in this alignment, and one SNP was not aligned by Gap4.

Detection of heterozygous sites

High quality read lengths (HQR) were defined as regions with aligned reads with minimum phred score of ≥ MVS, and phred scores ≥ 15 for all ten surrounding bases. Sites with ≥ 1 MVS SNP call and at least *N* (*N* ≥ 2) reads of HQR were used to estimate the proportion of heterozygous sites. A site qualified as heterozygous if the cumulative phred score from all reads covering that site was ≥ *H* (where H = 60, 70 or 80) for each of two alleles. Sites any evidence for with more than two alleles were excluded. For various values of *N* and *H*, estimates of the percentages of SNP locations that were heterozygous were: IT 4% - 6%, PFCLIN 7 - 10%.

Synonymous/non-synonymous rate ratios (dN/dS)

Pair wise alignments were generated using the predicted variant sequences for each comparison (IT, PFCLIN, *P. reichenowi*) to the reference for all protein-coding genes for all regions of *P. falciparum* genes covered by aligned reads from the 'unique' regions of the genome, excluding regions containing indels. The dN/dS was calculated from these

alignments for each gene with ≥ 300nt unique read coverage using the yn00 program of the PAML package[7], using the Yang and Nielsen algorithm. Only genes that contained at both synonymous change(s), and non-synonymous change(s) were included. Counts of the number of synonymous and non- synonymous changes in codons used the same method, all codons that were completely covered by reads were included.

McDonald-Kreitman tests (MK tests)

We carried out MK tests according to Ref [8]. The neutrality index (NI) was calculated from the equation[9]: NI = (Pn/Ps)/(Fn/Fs); Pn nonsynonymous polymorphisms, Ps synonymous polymorphisms, Fn nonsynonymous fixed differences, Fs synonymous fixed differences. The NI for groups of genes were calculated by concatenating all coding sequences for the group of genes. Significance was evaluated from concatenated gene sequences using Fisher's exact test.

Statistics

All statistics and plots were performed in R v1.11 (The R Foundation for Statistical Computing http://www.R-project.org). Tests for differences in evolutionary rate between groups of genes or correlations used only those genes included in the classification or detected in the expression study. For grouped data, if Kruskal-Wallis rank sum null hypothesis (that all groups contained were the same dN/dS distribution) was rejected ($P < 0.05$) then Mann-Whitney tests were performed on each group to examine whether the dN/dS distribution of the genes in the group differed from the dN/dS distribution of all other genes that were detected/classified in the expression study/classification.

Indel rates

Both PFCLIN to 3D7 and *P. reichenowi* to 3D7 comparisons showed more instances of "insertions" in 3D7 (or "deletions" in the other genomes), suggesting that the reference genome has more DNA. (t-test, number of insertions/10kb block vs. number of

deletions/10kb; PFCLIN t = -14.5455, df = 4129.046, P < 2.2×10^{-16}, *P. reichenowi* t = -4.2545, df = 3726.247, P = 2.147×10^{-5}).

The mean rate of the "insertion" in the reference genome was approximately +1 nt/kb (PFCLIN +1.04, *P. reichenowi* +0.94), these "insertion" rates were significantly greater than "deletion" rates. (t-tests, insertion length/10kb block vs deletion length/10kb; PFCLIN t = -11.3618, df = 4097.141, P < 2.2×10^{-16}, *P. reichenowi* t = -4.9538, df = 3698.472, P = 7.604×10^{-7}). We estmate that the 3D7 isolate has accumulated of the order of 20 kb of DNA in culture, based on the rate of +1 base/kb, over the genome of 23,267,177 bases.

Gene expression data

When analysing transcript expression data from Le Roch et al.[10] we observed that the number of developmental stages estimate was present in was significantly postively correlated with the total expression estimate (Spearman rank = 0.75, P < 2×10^{-16}), because low-expression genes were more likely to be detected only in one or a few stages. Because total expression is highly significantly correlated with dN/dS, correlations between dN/dS and number of expression stages were therefore not independent tests. To allow for correlation of evolutionary rate and number of stages to be examined independently of total expression, only those genes with total expression ≥ 200 were used to determine developmental stage, and number of stages expressed (number of stages and total expression were not correlated with this section of the data; Spearman rank, r = 0.028, P = 0.1378). This non-independence of measures was also observed for other expression data. To correct for this correlation in the protein expression data of Le Roch et al.[10], we used only genes with total protein > 100 for expression stages (then total protein counts and number of stages are not correlated: Spearman rank = 0.09344574, P = 0.1642). For the Young et al. data[11], we used only genes with at least ≥1400 minimal total signal, at which point the number of stages was not correlated with the total expression (Spearman rank r= -0.0004, P = 0.994).

The duration of expression (number of developmental stages expressed) still correlated with *P. reichenowi* dN/dS after these corrections; Le Roch et al. data[10] (Spearman rank correlations, protein = -0.37, P = 2.63×10^{-5}, transcript = -0.17, P = 2.7×10^{-12}). Young et al. data[11] (Spearman rank correlation, r = -0.23, P = 4.51×10^{-7}).

References

1. Macallan, D.C., Pocock, M., Robinson, G.T., Parker-Williams, J. & Bevan, D.H. Red cell exchange, erythrocytapheresis, in the treatment of malaria with high parasitaemia in returning travellers. *Trans R Soc Trop Med Hyg* **94**, 353-6 (2000).
2. Kyes, S., Pinches, R. & Newbold, C. A simple RNA analysis method shows var and rif multigene family expression patterns in Plasmodium falciparum. *Mol Biochem Parasitol* **105**, 311-5 (2000).
3. Dame, J.B. & McCutchan, T.F. Plasmodium falciparum: Hoechst dye 33258-CsCl ultracentrifugation for separating parasite and host DNAs. *Exp Parasitol* **64**, 264-6 (1987).
4. Gardner, M.J. et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511 (2002).
5. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-9 (2001).
6. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-6 (2000).
7. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-6 (1997).
8. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652-4 (1991).
9. Rand, D.M. & Kann, L.M. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* **13**, 735-48 (1996).
10. Le Roch, K.G. et al. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res* **14**, 2308-18 (2004).
11. Young, J.A. et al. The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* **143**, 67-79 (2005).